Tour d'Horizon du French QuestionBank : Construire un Corpus Arboré de Questions pour le Français

Djamé Seddah, Marie Candito

Tour d'Horizon du French QuestionBank : Construire un Corpus Arboré de Questions pour le Français

Djamé Seddah[1] Marie Candito[2]
(1) Inria (Almanach) & Université Paris Sorbonne
(2) CNRS (LLF) & Université Denis Diderot
djame.seddah@paris-sorbonne.fr,
marie.candito@linguist.univ-paris-diderot.fr

RÉSUMÉ

Nous présentons le French QuestionBank, un corpus arboré composé de 2600 questions annotées en dépendances et en constituants. Les deux tiers étant alignés avec le QuestionBank de l'anglais (Judge *et al.*, 2006), libre de droits, ce corpus saura prouver son utilité pour construire des systèmes d'analyse robuste. Nous discutons aussi des coûts de développement de tels corpus.

ABSTRACT

Building a Question Treebank for French : The French QuestionBank

We present the French QuestionBank, a treebank of 2600 questions, annotated with dependency phrase-based structures. Two thirds being aligned with the English QuestionBank (Judge *et al.*, 2006) and being freely available, this treebank will prove useful to build robust NLP systems. We also discuss the development costs of such ressources.

MOTS-CLÉS : corpus arborés, analyse syntaxique statistique, analyse hors domaine.

KEYWORDS: treebank, statistical parsing, out-of-domain parsing.

1 Introduction

[1] L'une des problématiques les plus récurrentes en analyse syntaxique statistique est la question de "l'analyse hors-domaine" (Foster *et al.*, 2007; McClosky *et al.*, 2010) ou "adaptation de domaine" d'un analyseur, c'est-à-dire la capacité pour un analyseur de traiter des phrases issues de corpus différant des corpus utilisés lors de la phase d'apprentissage. Les différences résultent typiquement de l'origine des corpus, on a souvent une situation où les corpus arborés servant à l'apprentissage sont issus de journaux, diffèrent drastiquement de phrases à analyser provenant de situations très diverses (questions d'utilisateurs, articles spécialisés etc...). Certains phénomènes présents dans les phrases à analyser peuvent être absents du corpus d'apprentissage, ou bien avoir une fréquence très différente. Sachant que la plupart des modèles d'apprentissage sont basés sur l'hypothèse que les données de test et d'entraînement ont les mêmes distributions, comment construire un analyseur capable de traiter n'importe quel type de texte à partir d'un corpus d'entraînement issu par définition d'un domaine restreint ?

On peut caractériser les divergences potentielles entre le corpus source et le domaine cible sur trois axes : (i) divergences lexicales, (ii) niveau de bruit, d'édition et (iii) divergences syntaxiques. Nous nous concentrons ici sur les divergences syntaxiques. L'intérêt des corpus annotés pour générer des

1. Cet article reprend partiellement et étend (Seddah & Candito, 2016).

modèles statistiques n'est plus à démontrer, cependant la très grande fragilité de ces modèles face aux variations de domaine motive l'enrichissement des corpus d'entraînement par d'autre sources de données suivant le même schéma d'annotation. Le cas du French Treebank (FTB, (Abeillé *et al.*, 2003)) est à cet égard symptomatique : étant constitué de phrases journalistiques (plus précisément un journal généraliste national, *Le Monde*), il ne contient que très peu d'interrogatives directes, rendant l'analyse automatique de telles formes difficile. Pour cette raison, nous avons développé un corpus recouvrant ces phénomènes, le French QuestionBank, un corpus arboré composé d'interrogatives syntaxiquement annotées, librement disponible sous 4 schémas d'annotation : deux en constituants (avec ou sans annotation des dépendances à longue distance via des chemins fonctionnels), deux en dépendances (surfaciques et en syntaxe profonde). Ce corpus présente en outre la particularité de contenir une partie importante de phrases alignées avec le QUESTIONBANK, son homologue pour l'anglais (Judge *et al.*, 2006).

2 Typologie des Phrases Interrogatives

Le but initial de ce travail était l'amélioration des performances des analyseurs statistiques dans un contexte de question-réponses, les interrogatives étant notoirement difficiles à analyser pour un analyseur statisitique en raison des structures non-canoniques qu'elles contiennent et de l'ordre des mots qu'elles impliquent (Dagnac, 2013; Beyssade, 2007). Plus précisément on peut distinguer les interrogatives à extraction du mot-*que* des interrogatives *in situ* à ordre des mots standard.

Les questions *In situ* se divisent en : **(i)** Celles contenant un syntagme interrogatif, i.e. un constituant contenant un marqueur interrogatif (adjectif, pronom ou adverbe) en position canonique (*Paul a mangé quel dessert ?, Alex a modifié quoi ?*) et **(ii)** les questions fermées pour lesquelles le status interrogatif est marqué soit uniquement par la prosodie (à l'écrit, par un point d'interrogation) (*Paul a déjà mangé ?*, soit par l'utilisation d'un clitique nominatif suivant le verbe fléchi. Le clitique est alors redondant ou pas avec un sujet pré-verbal non anaphorique (doublement du clitique : *Paul a-t-il déjà mangé ?* ou pas : *A-t-il déjà mangé ?* [2]).

Les interrogatives avec extraction présentent une structure et un ordre des mots bien différents, et sont donc susceptibles d'être plus difficiles à analyser automatiquement. On distingue :
cas (1) Extraction gauche avec sujet pré-verbal et doublement du clitique : *Quel dessert Paul a-t-il mangé ?*
cas (2) Extraction gauche avec inversion du sujet non-clitique : *Quel dessert a mangé Paul ?*
cas (3) Extraction gauche avec inversion du clitique sujet : *Quel dessert a-t-il mangé ?*
Seul le cas (2) peut apparaître dans une phrase ne formant pas une question (mais avec interrogative indirecte *Je sais quel dessert a mangé Paul*).

Les interrogatives mettant en jeu *(qu'/qui) est-ce que/qui* ont une syntaxe plus ou moins spécifique :
cas (4) Questions fermées de la forme *est-ce que + Phrase* : *Est-ce que Paul a déjà mangé ?* (formable de manière régulière à partir de *C'est que Paul a déjà mangé*)
cas (5) Forme en *qui/qu' est-ce que/qui + Phrase-à-trou* : *Qu'est-ce que Paul a mangé ?* (formable de manière régulière comme une question sur une clivée)
cas (6) Forme *qu' est-ce que + SN* : *Qu'est-ce que le platine ?* (spécifique)

2. Ce dernier cas correspond à l'inversion simple d'un clitique nominatif, et n'est donc pas *in situ*. Par commodité et lien avec le cas de doublement du clitique, nous comptons ce cas dans les *in situ* ci-après.

En résumé, les cas potentiellement problématiques pour un analyseur entraîné sur un corpus arboré classique sont principalement ceux contenant un mot-*que*. Les interrogatives *in situ* présentent un ordre canonique des mots mais la présence de mots-*que* peut poser des problèmes d'étiquetage. Les cas (1) à (6) sont plus problématiques encore en raison de l'ordre des mots non canonique.

	FTB-UC *(2007)*	FSMB *(2012)*	SEQUOIA *(2012)*	FQB *(-)*
# mots	350947	20584	69356	23236
# phrases	12351	1656	3204	2289
Long. moy. des phrases	28.41	12.42	21.64	10.15
# phrases avec synt. interrogatifs	210	61	85	1710
(%)	(1.68)	(3.68)	(2.65)	(74.7)
# Interrogatives avec extraction				
qu- cas 1	12	2	12	177
qu- cas 2	22	3	27	800
qu- cas 3	13	11	12	79
qu- cas 4	0	2	0	1
qu- cas 5	1	0	0	17
qu- cas 6	0	0	4	134
# questions *in situ*	172	54	30	502

TABLE 1 – Statistiques des corpus arborés du français les plus courants. Haut : Statistiques générales. Milieu : Nombre de phrases avec au moins un syntagme interrogatif (in situ ou pas). Bas : nombre d'interrogatives, classées selon la typologie de la section 2.

3 Les Interrogatives dans les Corpus Arborés du Français

Le *French treebank* (FTB) (Abeillé & Barrier, 2004) est le corpus arboré le plus utilisé pour l'entraînement d'analyseurs syntaxiques du français, pour des raisons historiques et liées à sa taille. D'autres sont apparus plus tard, en particulier le corpus SEQUOIA (Candito & Seddah, 2012b), un corpus de différentes origines (agence européenne du médicament, Europarl, wikipedia, *L'Est Républicain*), conçu au départ pour différer du genre du FTB (issu du journal *Le Monde*), et le FRENCH SOCIAL MEDIA BANK (FSMB) (Seddah *et al.*, 2012), représentatif des médias sociaux et des forums web.

Comme l'ont remarqué Judge *et al.* (2006) sur l'anglais, les interrogatives sont généralement sous-représentées dans les corpus arborés. Cette observation se vérifie aussi dans le cas du français pour lequel on recense seulement quelques centaines de syntagmes interrogatifs dans les corpus précédemment cités (Table 1).

4 Le French QuestionBank (FQB)

Origine des Données Les questions proviennent de plusieurs sources de corpus écrits [3] : (i) traduction vers le français du jeu de données test des campagnes TREC 8-11 [4], (ii) questions les plus fréquentes de plusieurs sites officiels (Trésor public, Pôle emploi, INSEE, Unesco), (iii) jeu de test de la campagne d'évaluation CLEF-03 sur des systèmes de Questions-Réponses (Magnini *et al.*, 2004) et (iv) questions issues des forums de MARMITON.ORG. Les trois premiers blocs sont correctement édités tandis que

3. Nous avons volontairement écarté les corpus oraux, qui nécessitent de gérer les disfluences.
4. Il s'agit de campagnes d'évaluation de systèmes de question-réponse (http ://www-rali.iro.umontreal.ca/rali/ ?q=node/9).

la section MARMITON, issue de contenus générés par l'utilisateur est extrêmement bruitée et présente des caractéristiques typiques aux médias sociaux qui rendent son analyse, tant en dépendance qu'en constituant, compliquée du fait des multiples ellipses qu'elle contient. De fait, nous n'abordons pas ici cette partie, étudiée en détail par Martínez Alonso *et al.* (2016).

Une collecte rendue difficile par le peu de données libres disponibles sur ce type de données, a conduit à un corpus relativement déséquilibré comparé au QuestionBank (QB, (Judge *et al.*, 2006)) comme le montrent les différences de tailles des composantes de notre corpus (cf. Table 2). Notons que la partie TREC du French QuestionBank (FQB) est alignée avec les 1893 premières phrases du QB. L'utilisation conjointe de ces deux ressources pourrait être utilisée pour l'évaluation de systèmes de traduction orientés syntaxe voire pour l'amorçage de tels systèmes.

Schéma d'annotation

Constituants Afin d'obtenir des données annotées compatibles avec des analyseurs entraînés sur le FTB, nous avons utilisé le schéma d'annotation de celui-ci comme base de travail et suivi autant que possible son guide d'annotation pour la morphologie, la structure de constituants et les annotations fonctionnelles (Abeillé *et al.*, 2003; Abeillé & Barrier, 2004). Plus précisément, nous sommes partis des modifications légères apportées par Candito & Crabbé (2009), ci-après FTB-UC, et avons étendu le guide afin de couvrir les spécificités syntaxiques des interrogatives.

SOURCE	# DE PHRASES
TREC 08-11	1893
Faq GVT/NGOs	196
CLEF03	200
sub-total	*2289*
Web (Marmiton)	285

TABLE 2 – Source des données du FQB.

Labels fonctionnels et dépendances longues distances Concernant les étiquettes fonctionnelles, nous avons utilisé un label additionnel DIS, propre aux composants disloqués , introduit par (Abeillé & Crabbé, 2013). Ceux-ci apparaissent en début ou en fin de clause et sont coréférents d'un clitique, apparaissant sur le verbe. Ces fonctions apparaissent parfois dans des phrases déclaratives (*Paul les/OBJ a mangées, les fraises/DIS*) mais ont été massivement utilisées dans le FQB pour les interrogatives en *Qu'est-ce que SN*, dont on trouvera un exemple d'annotation en Figure 1. Celles-ci relèvent du cas 6, présenté section 2.

Afin de préparer la conversion en dépendances surfaciques et profondes (cf. section suivante), nous avons aussi annoté les dépendances longue distance, à l'aide de chemins fonctionnels (Schluter & van Genabith, 2008; Candito & Seddah, 2012a), de manière à récupérer le gouverneur correct lors de la conversion [5].

Dépendances surfaciques et profondes Une couche d'analyse en dépendance surfacique est ensuite obtenue en utilisant la conversion présentée dans (Candito *et al.*, 2010)et légèrement modifiée pour prendre en compte les extensions présentées plus haut (cf. Figure 2 arcs noirs et arcs rouges). A partir de cette sortie, nous avons ajouté une surcouche d'annotation en syntaxe profonde (cf. Figure 2 arcs noirs et arcs bleus) respectant le schéma d'annotation

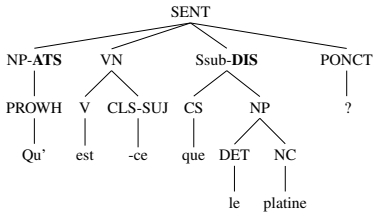

FIGURE 1 – Analyse avec disloquée

5. Par exemple la conversion standard de *Quelles démarches dois-je effectuer ?* donne *démarches* dépendant de *dois*, l'annotation manuelle d'un chemin fonctionnel permet de rattacher *démarches* à *effectuer*. A noter que cela donne ici de la non projectivité.

DEEP SEQUOIA (Candito *et al.*, 2014; Perrier *et al.*, 2014) et suivant le protocole conçu pour la conversion du FTB vers ce même schéma (Ribeyre *et al.*, 2014). Cette conversion suppose une annotation manuelle préalable du statut des clitiques "se" et des "il" explétifs. Elle produit des graphes de dépendances. Les dépendances profondes neutralisent en particulier les changements de diathèse, en explicitant si besoin les fonctions "canoniques" en plus des fonctions "finales" (cf. les doubles labels $final$:$canonique$ dans la figure 2).

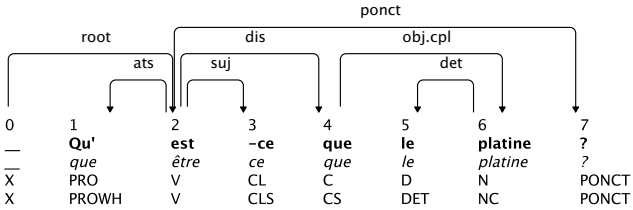

FIGURE 2 – Analyse en dépendances de surface

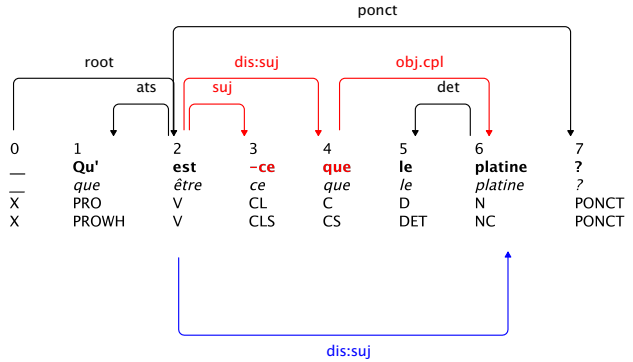

FIGURE 3 – Analyse en dépendances "de surface" (arcs noirs et arcs rouges), et analyse en dépendances profondes (arcs noirs et arcs bleus). Les labels de la forme x :y signifient que la fonction finale est x mais la fonction canonique est y.

Méthodologie d'Annotation et Evaluation Nous avons suivi le même protocole d'annotation que Candito & Seddah (2012b) : On commence par doublement annoter puis adjudiquer l'annotation en parties du discours. Ensuite, afin de réduire les biais d'annotation, deux annotateurs travaillent sur les sorties de deux analyseurs syntaxiques prenant en entrée les questions avec parties du discours validées (l'analyseur de Berkeley (Petrov *et al.*, 2006) et l'analyseur *first-phase* de Charniak (2000)). Les sorties corrigées sont ensuite adjudiquées. On réalise ensuite l'annotation des étiquettes fonctionnelles (fonctions grammaticales des dépendants de verbes) selon le même protocole. On annote ensuite les dépendances longue distance, le statut des "se" et des "il", et on applique enfin la conversion automatique en arbres de dépendances de surface, puis en graphes de dépendances profondes.

Pour estimer la qualité de l'annotation, et suivant en celà la méthodologie de Candito & Seddah (2012b), nous avons calculé l'accord inter-annotateurs en utilisant la métrique F-score de PARSEVAL [6] entre deux annotations en constituants avec étiquettes fonctionnelles. Les scores présentés Table 3 sont légèrement plus élevés que ceux rapportés lors de précédentes campagnes d'annotation de corpus arboré hors-domaine (Candito & Seddah, 2012b; Seddah *et al.*, 2012). Ceci est sans doute dû à la faible longueur moyenne des phrases du FQB et au fait que ses annotateurs étaient déjà entraînés à cette tâche et sur ce schéma d'annotation.

A vs B	A vs Référence	B vs Référence
97.54	95.72	97.21

TABLE 3 – Accord inter-annotateurs (A vs B) et entre chaque annotateur et la référence adjudiquée, évalué par un F-score du parenthésage étiqueté.

4.1 Les coûts cachés de l'annotation de corpus arborés

Comme nous le savons tous, la création de données est une tâche gratifiante, extrêmement utile pour l'étude, la constitution et l'évaluation de modèles linguistiquement motivés. Cependant, ce processus est très gourmand en temps et d'une façon générale, assez coûteux (Böhmová *et al.*, 2003; Schneider, 2015) même si des solutions basées sur la myriadisation ou sur le principe des jeux avec objectifs sont en train d'émerger, et, pour certaines, sont arrivées à maturité (Guillaume *et al.*, 2016). Le jeu de données que nous avons présenté dans cet article fait partie d'un processus initié il y a 5 ans lorsque nous avons été confrontés à l'absence de données hors-domaine annotées en syntaxe pour le français. [7]

Il est important de préciser qu'une telle tâche a été rendue possible par l'expérience que nous avons acquise durant ces années et parce que nous nous sommes appuyés sur une équipe hautement qualifiée d'annotateurs. Cette formation était le point le plus important en termes de coûts et devait être étendue pour chaque nouveau domaine ou pour chaque nouvelle couche d'annotation (de la syntaxe de surface à la syntaxe profonde par exemple, voir (Candito *et al.*, 2014) pour plus de détails).

La table 4 présente les coûts des efforts d'annotations de corpus arborés (principalement dirigés par les deux co-auteurs de ce papiers) qui ont été portés par l'équipe Alpage depuis 2011 et qui ont conduit à la publication de plusieurs corpus arborés hors-domaine et de contenus générés par l'utilisateur (UGC). Ces sommes ne couvrent pas les coûts des personnels permanents impliqués dans le développement des schémas d'annotation, des annotations préliminaires, des outils de pré-annotation, de ceux de détections et corrections d'erreurs post-annotation et bien sûr du coût de formation et de supervisions des annotateurs. [8] En considérant que nous avons annoté 4 couches d'annotation différentes pour près de 7000 phrases hors-domaines et deux couches pour 3700 phrases d'UGC, on obtient un coût moyen de 3 euros par phrase et par couche, comparable aux coûts de développement de l'English Web Treebank (Bies *et al.*, 2012) indiqués par Ryan Mc Donald (P.C.) et confirmés ensuite par

6. Il s'agit de la moyenne harmonique du rappel et de la précision sur les constituants étiquetés. Un constituant est compté comme correct s'il est du même type et contient exactement les mêmes mots que dans l'autre annotation, considérée comme référence. À noter qu'inverser ce qui est considéré comme référence dans le calcul inverse simplement précision et rappel, sans impact sur le F-score.

7. Cet effort est par ailleurs toujours en cours dans le cadre du projet ANR ParSiTi, visant à produire entre autres des corpus arborés parallèles de contenus générés par l'utilisateur multilingues.

8. On pourrait approximer les coûts d'environnement en prenant comme base de calcul les taux type ANR (8%) ou H2020 (25%).

Fernando Perreira. [9] Notons que contrairement à d'autres initiatives moins coûteuses qui se sont concentrées sur la création de corpus de référence (*gold standard*) et le développement rapide de corpus d'entraînement, notre objectif est aussi de proposer un instantané linguistique d'un domaine spécifique à un instant donné, utile en tant que ressource d'exploration linguistique.

	Départ	Taille *phrases*	Morph.	Const.	Dep. *mois-hommes*	Deep Synt.[5]	Cost *euros*
Sequoia[1]	2011	3200	2	9	1	6	59k
FSMB 1[2]	2012	1700	1	2	-	-	13k
FSMB 2[2]	2014	2000	2	4	-	-	20k
FQB[3]	2014	2600	2	4	1	4	36k
LoL[4]	2015	450	3	-	-	-	3k
Minecraft[4]	2016	230	0.5	-	-	-	2k
		10180			41.5		133k

[1] *:(Candito & Seddah, 2012b),* [2] *:(Seddah* et al.*, 2012),*[3] *:(Seddah & Candito, 2016)*
[4] *:(Martínez Alonso* et al.*, 2016),*[5] *:(Candito* et al.*, 2014)*

TABLE 4 – Coût des campagnes d'annotation à Alpage. *Morph. : annotations morpho-syntactiques , Const. : annotation en constituants, Dep. : conversion en dépendances, Deep Synt : annotation en syntaxe profonde*

Conclusion

Nous avons présenté le French QuestionBank, le premier treebank de questions en dehors de l'anglais, et partiellement alignée avec son homologue. Seddah & Candito (2016) ont démontré son utilité pour l'analyse d'un genre de texte sous-représenté dans les jeux de données annotés en syntaxe actuellement disponibles, sans nuire aux performance *in-domain* des analyseurs. L'inclusion du FQB dans tout modèle permet donc d'obtenir des analyseurs moins sensibles aux divergences de domaines et de ce fait justifie son développement. Nous avons abordé la question des coûts liés au processus d'annotation, un point rarement évoqué dans la littérature.

Le FQB est distribué sous licence LGPL-LR à l'adresse suivante `http://alpage.inria.fr/Treebanks/FQB/`.

Remerciement

Nous remercions nos relecteurs anonymes pour leurs commentaires (certains seulement pris en compte par manque de place). Nous remercions également Benoit Crabbé et Corentin Ribeyre pour leurs commentaires sur une version précédente de ce travail. Les auteurs ont bénéficié du soutien du Programme "Investissement d'avenir" de l'Agence Nationale de la Recherche ANR-10-LABX-0083 (Labex EFL) et du projet ANR SoSweet (ANR-15-CE38-0011-01).

Références

ABEILLÉ A. & BARRIER N. (2004). Enriching a french treebank. In *Proc. of LREC'04*, Lisbon, Portugal.

ABEILLÉ A., CLÉMENT L. & TOUSSENEL F. (2003). *Building a Treebank for French*, In *Treebanks*. Kluwer : Dordrecht.

ABEILLÉ A. & CRABBÉ B. (2013). Vers un treebank du français parlé. In *In Proc. of TALN 2013*, Les Sables d'Olonnes, France.

9. `https://twitter.com/earnmyturns/status/794741252285566980`

BEYSSADE C. (2007). La structure de l'information dans les questions : quelques remarques sur la diversité des formes interrogatives en fran ç ais. *LINX*, p. 173–193.

BIES A., MOTT J., WARNER C. & KULICK S. (2012). *English Web Treebank*. Rapport interne, Linguistic Data Consortium,, Philadelphia, PA, USA.

BÖHMOVÁ A., HAJIČ J., HAJIČOVÁ E. & HLADKÁ B. (2003). The prague dependency treebank. In *Treebanks*, p. 103–127. Springer.

CANDITO M. & CRABBÉ B. (2009). Improving generative statistical parsing with semi-supervised word clustering. In *Proc. of IWPT'09*, Paris, France.

CANDITO M., CRABBÉ B. & DENIS P. (2010). Statistical french dependency parsing : Treebank conversion and first results. In *Proc. of LREC*.

CANDITO M., PERRIER G., GUILLAUME B., RIBEYRE C., FORT K., SEDDAH D. & DE LA CLERGERIE É. V. (2014). Deep syntax annotation of the sequoia french treebank. In *International Conference on Language Resources and Evaluation (LREC)*.

CANDITO M. & SEDDAH D. (2012a). Effectively long-distance dependencies in French : annotation and parsing evaluation. In *Proc. of TLT 11*.

CANDITO M. & SEDDAH D. (2012b). Le corpus sequoia : annotation syntaxique et exploitation pour l'adaptation d'analyseur par pont lexical. In *In Proceedings of Traitement Automatique des Langues Naturelles (TALN 2012)*, Grenoble, France.

CHARNIAK E. (2000). A maximum entropy inspired parser. In *Proceedings of the First Annual Meeting of the North American Chapter of the Association for Computational Linguistics (NAACL 2000)*, p. 132–139, Seattle, WA.

DAGNAC A. (2013). La variation des interrogatives en français. document préparatoire (texte provisoire) pour contribution à la GGF (Abeillé, A., Godard, G. et A. Delaveau, éds, à paraître en 2017).

FOSTER J., WAGNER J., SEDDAH D. & VAN GENABITH J. (2007). Adapting wsj-trained parsers to the british national corpus using in-domain self-training. In *Proceedings of the Tenth IWPT*, p. 33–35.

GUILLAUME B., FORT K. & LEFEBVRE N. (2016). Crowdsourcing complex language resources : Playing to annotate dependency syntax. In *Proceedings of the International Conference on Computational Linguistics (COLING)*, Osaka (Japon).

JUDGE J., CAHILL A. & VAN GENABITH J. (2006). QuestionBank : Creating a Corpus of Parse-Annotated Questions. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Anual Meeting of the Association for Computational Linguistics (COLING-ACL 2006)*, p. 497–504, Sydney, Australia.

MAGNINI B., ROMAGNOLI S., VALLIN A., HERRERA J., PEÑAS A., PEINADO V., VERDEJO F. & DE RIJKE M. (2004). Creating the disequa corpus : a test set for multilingual question answering. In *Comparative Evaluation of Multilingual Information Access Systems*, p. 487–500. Springer.

MARTÍNEZ ALONSO H., SEDDAH D. & SAGOT B. (2016). From noisy questions to minecraft texts : Annotation challenges in extreme syntax scenario. In *Proceedings of the 2nd Workshop on Noisy User-generated Text (WNUT)*, p. 13–23, Osaka, Japan : The COLING 2016 Organizing Committee.

MCCLOSKY D., CHARNIAK E. & JOHNSON M. (2010). Automatic domain adaptation for parsing. In *Human Language Technologies : The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, p. 28–36 : Association for Computational Linguistics.

PERRIER G., CANDITO M., GUILLAUME B., RIBEYRE C., FORT K. & SEDDAH D. (2014). Un schéma d'annotation en dépendances syntaxiques profondes pour le français. In *TALN-Traitement Automatique des Langues Naturelles*, p. 574–579.

PETROV S., BARRETT L., THIBAUX R. & KLEIN D. (2006). Learning accurate, compact, and interpretable tree annotation. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*, Sydney, Australia.

RIBEYRE C., CANDITO M. & SEDDAH D. (2014). Semi-automatic deep syntactic annotations of the french treebank. In *Proceedings of the 13th International Workshop on Treebanks and Linguistic Theories (TLT13)*, Tübingen, Germany.

SCHLUTER N. & VAN GENABITH J. (2008). Treebank-based acquisition of lfg parsing resources for french. In E. L. R. A. (ELRA), Ed., *Proceedings of the Sixth International Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco.

SCHNEIDER N. (2015). What I've learned about annotating informal text (and why you shouldn't take my word for it). In *The 9th Linguistic Annotation Workshop held in conjuncion with NAACL 2015*, p. 152.

SEDDAH D. & CANDITO M. (2016). Hard time parsing questions : Building a questionbank for french. In *in Proceedings of LREC 2016*.

SEDDAH D., SAGOT B., CANDITO M., MOUILLERON V. & COMBET V. (2012). The french social media bank : a treebank of noisy user generated content. In *Proceedings of COLING'12*, Mumbai, India.